MésoLUM, mésocentre HPC multi-branches

Georges Raseev et Philippe Dos Santos

Fédération de Recherche LUmière MATière (FR LUMAT), Université Paris-Sud, Orsay

INTRODUCTION

La modélisation des phénomènes en physique, chimie et biologie, particulièrement dans le cas de l'interaction laser-matière, occupe une place importante dans la recherche scientifique. Pour effectuer cette modélisation les moyens de calcul ont évolué vers l'utilisation répandue de grappes de calcul HPC (High Performance Computing). Gérer et pérenniser de tels outils de calcul passe par la mise en place de structures appelées mésocentres disposant d'informaticiens qualifiés formés aux techniques du HPC. Parmi les mésocentres de l'Université Paris-Saclay le Mésocentre LUmière Matière (acronyme MésoLUM), plate-forme de la Fédération de Recherche LUmière MATière (FR LUMAT) (site internet http://www.mesolum.lumat.u-psud.fr), se distingue par une architecture multi-branches. Cette architecture en branches permet la continuité de service, c'est à dire que le mésocentre fonctionne sans interruption aucune 365 jours par an, même si une branche est arrêtée. Cet article présente l'organisation en branches du mésocentre MésoLUM, qui pourrait avoir un rôle structurant dans l'organisation des moyens de calcul de l'Université Paris-Saclay. En plus de la continuité de service d'autres utilisations de cette architecture sont possibles par exemple en spécialisant une branche avec des GPU ou des Xeon-phi.

L'ARCHITECTURE EN BRANCHES ET LA CONTINUITE DE SERVICE

MésoLUM, mésocentre et plate-forme de la FR LUMAT, mutualise plusieurs grappes de calcul HPC. Une grappe HPC est constituée de :

i) plusieurs nœuds de calcul rapides renouvelés en partie tous les 12 mois. Ce renouvellement est indispensable puisque la puissance de calcul des processeurs augmente rapidement,

ii) un réseau rapide à faible latence permettant des calculs rapides en parallèle sur plusieurs nœuds. Ce réseau est nécessaire pour diminuer le temps global d'exécution des programmes,

iii) un serveur de fichiers parallèle pour le HPC. Les performances de ce type de serveur de fichiers sont constantes en lecture/écritures (I/O) quel que soit le nombre d'utilisateurs.

Le mésocentre MésoLUM est essentiellement utilisé pour les modélisations numériques effectuées par les chercheurs des laboratoires de la FR LUMAT (LUmière MATière, Fédération de Recherche du CNRS) avec des programmes développés en interne par les chercheurs et des programmes externes open source ou commerciaux (comme par exemple Gaussian, VASP, Turbomol, Abinit). Afin de répondre à ces besoins, la puissance de calcul de MésoLUM est passée de 1 TFLOPS* CPU** en 2008 à 15 TFLOPS* CPU** et 4 TFLOPS* GPU*** fin 2015.

---

* TFLOPS : 10^{12} Floating point Operations Per Second
** CPU : Central Processing Unit ou processeur central d'un ordinateur
*** GPU : Graphics Processing Unit ou processeur de traitement d'une carte graphique

---

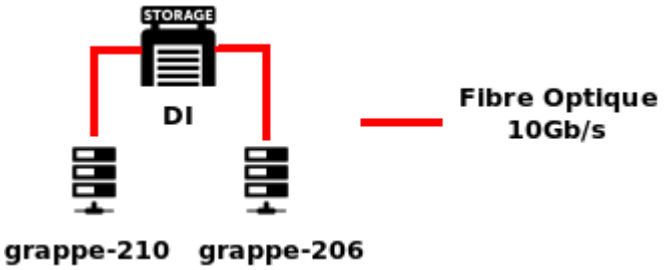
Par rapport à la définition ci-dessus, MésoLUM présente une structure en branches assurant la continuité de service. Pour assurer cette continuité de service, permettre l'augmentation de la puissance de calcul et mutualiser l'infrastructure, courant 2015 MésoLUM a été divisée en trois éléments hébergés dans trois salles informatiques indépendantes.

Ces éléments sont (voir figure) :

i) le serveur de fichiers parallèle localisé dans une salle sécurisée de la Direction Informatique (DI) de l'Université Paris-Sud. Il est relié à chacune des branches par une liaison fibre optique à 10 Gb/s dédiée aux transferts de fichiers,

ii) la branche appelée grappe-210 localisée dans une salle informatique sécurisée de l'ISMO. Elle est composée d'un nœud maître, de nœuds de calcul, d'un système de stockage parallèle dédié aux fichiers temporaires de grande taille (>10To) et d'un réseau rapide à faible latence connectant les nœuds de calcul,

iii) la branche appelée grappe-206 localisée dans la salle informatique VirtualData. Elle a la même structure que la branche grappe-210.



Depuis décembre 2015, deux branches de MésoLUM sont en service. L'accès aux branches est transparent pour les utilisateurs. En effet, un point d'entrée unique, l'adresse IP, est partagé par les branches via un mécanisme d'adresse IP Virtuelle. Indépendamment de la branche sur laquelle il travaille, l'utilisateur dispose d'une vue unifiée des nœuds de calcul de toutes les branches. En fonction des ressources disponibles, les calculs sont distribués automatiquement entre les nœuds des branches. La continuité de service implique que l'utilisateur travaille de la même façon si une branche est indisponible.

C'est l'organisation décrite ci-dessus en trois éléments qui permet la continuité de service de MésoLUM. Par exemple, pendant le déménagement de l'ISMO qui se déroulera en 2017, la branche grappe-206 reste disponible tandis que la branche grappe-210, à l'arrêt, sera déplacée dans la salle informatique du nouveau bâtiment de l'ISMO.

Les autres aspects de l'organisation du mésocentre MésoLUM sont :

- **Maintenance et évolution** : notre politique de maintenance et d'évolution du matériel informatique tient compte de l'évolution très rapide de la puissance de calcul des processeurs. Nous achetons des nœuds disposant d'une garantie de 3 ans. Au bout de cette période, le nœud devient obsolète et n'est plus utilisé ;
- **Temps de restitution** : le temps de restitution, temps entre la soumission et la restitution d'un calcul, est optimisé sur MésoLUM. Pour profiter pleinement des performances du HPC, un calcul est affecté à une branche et une seule ;
- **Support** : le support aux utilisateurs consiste d'une part à installer et à valider les programmes de calcul et d'autre part à adapter ces programmes au calcul parallèle (MPI essentiellement) ;
- **Financement** : Sans budget récurrent d'investissement, pour le renouvellement partiel des nœuds, l'équipement mutualisé MésoLUM est donc financé par les projets de recherche soumis soient par l'équipe du MésoLUM soient soumis par les membres de la FR LUMAT. Dans ce dernier cas, le chercheur ayant obtenu un financement pour des nœuds de calcul aura la priorité pour l'utilisation de ces nouvelles ressources, mais pas l'exclusivité. Ainsi, lorsque ces nœuds de calcul ne sont pas utilisés par ce chercheur, ils sont accessibles à l'ensemble de la communauté ;

CONCLUSION

La plate-forme MésoLUM de la FR LUMAT est un mésocentre mutualisé permettant à plus de 60 chercheurs et enseignants-chercheurs de faire des calculs théoriques et des modélisations numériques en support à leurs projets scientifiques.

Pour assurer une continuité de service, en décembre 2015 une seconde branche de MésoLUM a été mise en service. Précisément, cette continuité de service a été réalisée en localisant les trois éléments de MésoLUM reliés par fibre optique dans des salles informatiques sécurisées et indépendantes. La continuité de service permet au MésoLUM de fonctionner même lors de l'indisponibilité d'une branche.

L'architecture actuelle en deux branches peut immédiatement être étendue à plusieurs branches. Cette architecture est une innovation au niveau d'un mésocentre HPC et elle peut permettre de structurer et de mutualiser les moyens de calcul de l'Université Paris-Saclay, par exemple en distribuant la charge de calcul entre plusieurs grappes. De même, il est possible de spécialiser une branche (GPU, Xeon Phi, etc…) et de la rendre accessible à une communauté de chercheurs beaucoup plus large. Dans le cadre du projet Exascale Computing @ Paris Saclay (ECPS) déposé à l'appel des Initiatives de Recherche Stratégiques (IRS) de l'Université Paris-Saclay porté par la Maison de la Simulation, nous avons proposé de partager notre expérience d'une grappe de calcul à deux branches pour l'ensemble de la communauté des mésocentres de l'Université Paris-Saclay.